/

ECE 538 - Advanced Computer Architecture
Mid-Term Exam (Individual)

90

Date: 10/27/2021 **Time:** 2:00PM – 3:30PM

Test Guidelines:

- Open-book, Open-note, No Internet.
- Calculators are permitted, but no communication devices of any kind are allowed.
- Show your answers in the space provided for them. Write neatly and be well organized.
- The test is due exactly at the end of the class period. Each question is marked with its number of points, use your time wisely.

Suggestion:

- Whenever possible, show your work and your thought process. This will make it easier for us to give you partial credit.

Please sign your name below to indicate that you have read the above and have followed all the rules in the Academic Integrity according to the UNM student handbook.

Name: David Kirby

ID: 101652098

Score:

1. (20 points) There is a designed 5-stage, pipelined processor and synthesized it for a 45nm process technology node with a target clock rate of 1GHz. During power analysis, the processor is at the target clock rate with a supply voltage of 1.0V. This processor draws 70mW of dynamic power and 10mW of static power. Answer following questions by considering power and energy trade-offs:

   (a) (5 points) If a cryptographic operation takes 0.5 second to complete on the processor, what is the energy per operation at the target clock rate assuming back-to-back encryption operations?

5

$Power_{total} = Power_{dynamic} + Power_{static}$

$= 70mW + 10mW$

$= 80mW$

$Energy\ per\ operation = Power_{total} \times time\ to\ complete\ operation$

$= 80mW \times 0.5s$

$= 0.04\ J/op$ ✓

   (b) (5 points) For common cryptographic operations on the processor, if you were to slow the clock down to 500MHz without adjusting the voltage, what would be the energy per operation? What would be the overall power draw?

5

$Power'_{dynamic} = Power_{dynamic} \times \frac{new\ clk\ rate}{original\ clk\ rate}$ $Power_{static}$ stay same!

$= 70mW \times \frac{500MHz}{1GHz} \rightarrow \frac{1}{2}$

$= 35mW$

$Power'_{total} = Power'_{dynamic} + Power_{static}$

$= 35mW + 10mW$

$= 45mW$

$Energy\ per\ operation = Power'_{total} \times time\ to\ complete\ operation$

$= 45mW \times 0.5s$

$= 0.0225\ J/op$ ✓

   (c) (5 points) If safely drop the voltage to 0.5V when operating at a 500MHz clock, recalculate the power draw and energy per operation. Assume the leakage current remains the same. Please indicate the name of this energy saving technique with brief explanation.

5

$Power'_{dynamic} = P_{dynamic} \times \frac{new\ clk\ rate}{original\ clk\ rate} \times \left(\frac{new\ voltage}{old\ voltage}\right)^2$

$= 70mW \times \left(\frac{500MHz}{1GHz}\right) \times \left(\frac{0.5V}{1.0V}\right)^2$

$= 8.75mW$

$Power'_{static} = Power_{static} \times \frac{new\ volt}{old\ volt}$

$= 10mW \times \left(\frac{0.5V}{1V}\right)$

$= 5mW$

$Power'_{total} = Power'_{dynamic} + Power'_{static} = 13.75mW$

$Energy\ per\ operation = P_{total} \times time\ to\ complete = 13.75mW \times 0.5s = 0.006875\ J/op$

This is the dynamic voltage frequency scaling technique

   (d) (5 points) Assuming your system performs one operation every second and gates the clock off in between when not performing a cryptographic operation, what would be the energy per operation? Also assume the original 1GHz clock rate and 1.0V supply voltage.

?

2. **(10 points)** There is one cluster supported by the Center for Advanced Research Computing (CARC) at UNM. Assume the cluster has 500 computers, each of them with a MTTF of 25 days, and the failures follow an exponential distribution and are independent.

   (a) (5 credits) If 1/5 of the computers fail, the cluster is considered to fail. What is the MTTF of the cluster? Under this failure model, does adding more computers increase the MTTF on the cluster? Why?

$$MTTF = \frac{25 \text{ days}}{500 \text{ computers}} \times \frac{1}{5} \times 500 = 5 \text{ days}$$

5 Under this failure model, MTTF is not dependent on the number of computers and therefore adding more would not have an effect on the MTTF of the cluster since we are monitoring a $\frac{1}{5}$ ratio

   (b) (5 credits) For the same amount of money, one could buy 800 computers, each with MTTF of 20 days. Assume that the cluster (implementation with either 800 less reliable computers or 500 original computers) is considered to fail if a single computer fails. Repairing the less reliable cluster configuration is 10% less expensive. Which cluster would be better?

$$MTTF = \frac{25}{500} = \frac{1}{20} \text{ days} = 1.2 \text{ hours}$$

$$MTTF' = \frac{20}{800} = \frac{1}{40} \text{ days} = 0.6 \text{ hours}$$

5 Failure rate is doubled and even with 10% less expensive equipment, I would recommend the older, more reliable cluster

3. (15 points) Define and briefly describe the types of cache misses (also called the three C's) discussed in lecture. List other type(s) of cache misses as you know.

1) Compulsory: these are initial misses due to an empty (cold) cache

15 2) Conflict: these are the misses due to a rigid block placement strategy (i.e., low associativity)

3) capacity: These are misses due to the cache being too small to hold the entire working set of data and instructions

others) coherency: misses due to the cache coherence protocol used for sharing memory amongst processors

4. (20 points) Suppose that in 1000 memory references there are 25 misses in the L1 cache and 12 misses in the L2 cache.

(a) (5 points) Calculate the local miss rate and the global miss rate for the L1 and L2 caches.

L1 Local miss rate = $\frac{\#\text{ misses}}{\#\text{ accesses}} = \frac{25}{1000} = \frac{1}{40} = 0.025$ or 2.5%

Global miss rate = local miss rate for L1 = $\frac{25}{1000} = \frac{1}{40} = 0.025$ or 2.5%

4

L2 Local miss rate = $\frac{\#\text{ misses}}{\#\text{ accesses}} = \frac{12}{1000} = \frac{3}{250} = 0.012$ or 1.2%

Global Miss rate = $\text{Local}_{L1} \times \text{Local}_{L2} = 0.025 \times 0.012 = 0.0003$ X or 0.03%

(b) (5 points) Assume the miss penalty from L2 cache to main memory is 45 clock cycles, the hit time of L2 cache is 10 clock cycles, and the hit time of L1 is 1 clock cycle. Calculate the average memory access time.

$AMAT = \text{Hit Time}_{L1} + \text{Miss rate}_{L1} \times (\text{hit time}_{L2} + \text{missrate}_{L2} \times \text{penalty}_{L2})$

$= 1cc + 0.025 \times (10cc + 0.012 \times 45cc)$

3

$= 1.025 \times (10.54cc)$

$= 10.8035\,cc$ X

(c) (5 points) Assume this is an in order pipeline. If there are 1.25 memory references per instruction, what are the average stall cycles per instruction?

$\frac{\text{stalls}}{\text{instr.}} = \frac{\text{misses}_{L1}}{\text{instr}} \times \text{hit time}_{L2} + \frac{\text{misses}_{L2}}{\text{instr}} \times \text{miss penalty}_{L2}$

3

$= \frac{0.025}{1.25} \times 10cc + \frac{0.012}{1.25} \times 45cc$

$= 0.2 + 0.432 = 0.432 \frac{cc}{instr}$ X

(d) (5 points) Briefly describe the principles of the temporal and spacial locality. Explain how they relate to caching.

Spatial locality states that if a data location is referenced, there is a high liklihood that nearby locations will be referenced as well. Temporal deals with time and states that if a data location is referenced, there is a high liklihood that it will be referenced again. Instruction cache uses lots of spatial and temporal locality compared to data access, and helps to speed up the memory hierarchy.

5

5. (15 points) List the techniques to reduce the cache miss rate, miss penalty, and hit time with brief discussion (e.g., the effect of too large block size).

15

Reduce miss rate by:

- larger block sizes: spatial locality but large block sizes suffer from conflict misses
- larger caches: but longer hit time
- higher associativity
- multilevel caches

Reduce miss penalty by:

- giving priority to read misses over writes
- critical word first, merging write buffers, and prefetching

Reduce hit time by:

- avoiding address translation during indexing and small, simple 1st level cache.

6. (20 points) Describe how write buffers are used to improve the performance of the memory hierarchy. Indicate Where the write buffers exist in the memory hierarchy. Use diagram to help with the description.

20

Write buffers improve performance by reducing cache misses and reducing unnecessary stalling on writes, especially for write-through.

Write buffers can exist essentially any where in the memory hierarchy, for example between the L1 and L2 caches as shown below. Also between memory and L2.